Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 2518 – 2525

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Using CrowdFlower to study the relationship between self-reported violations and traffic accidents

J.C.F. de Winter*, M. Kyriakidis, D. Dodou, R. Happee

Department BioMechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

Abstract

Crowdsourcing is a promising approach for Human Factors survey research. We explored the use of a relatively new crowdsourcing platform called CrowdFlower. Our survey focused on the relationship between self-reported traffic accidents and violations measured with the Driver Behaviour Questionnaire (DBQ). We obtained 1,862 responses within 20 hours at a cost of $247. The demographic correlates of DBQ violations were consistent with those of traditionally recruited samples. The correlation between DBQ violations and self-reported accidents was ρ = .28. Self-reported accidents at the national level (*N* = 18 countries) correlated strongly (ρ = .68/.79) with accident statistics published by the World Health Organization. Large international differences were observed, with horn honking being relatively common in India and Indonesia and speeding being common in some Western countries. We conclude that CrowdFlower is an efficient tool for conducting international surveys.

Peer-review under responsibility of AHFE Conference

Keywords: Crowdsourcing; CrowdFlower; Driver Behaviour Questionnaire; traffic violations; traffic accidents

1. Introduction

1.1. Crowdsourcing as an alternative to traditional survey methods

Surveys are widely used to collect information from the general public. Traditional surveys attempt to recruit representative samples via mail, telephone, or face-to-face interviews. These survey methods are usually expensive and laborious, resulting not only in a long time to publication, but also in relatively small sample sizes. Recent

* Corresponding author. Tel.: +31-15-2786794.
E-mail address: j.c.f.dewinter@tudelft.nl

Peer-review under responsibility of AHFE Conference
doi:10.1016/j.promfg.2015.07.514

literature has indicated that surveying large and representative samples via the Internet is possible (cf. the Gosling-Potter Internet Personality Project, containing data of more than 3 million respondents [1,2]). Unfortunately, distributing and promoting an Internet survey may require substantial investments. Crowdsourcing refers to obtaining services from a dedicated online community. Opposite to 'traditional' Internet surveys, crowdsourcing-based surveys are usually conducted in return of a small fee. Advantages of crowdsourcing compared to physical recruitment are completion speed, population diversity, anonymity, reduced social desirability, and low costs [3].

1.2. Amazon Mechanical Turk (MTurk): The dominant crowdsourcing platform for research purposes

Researchers have come to appreciate crowdsourcing as a research tool. Launched in November 2005, Amazon Mechanical Turk (MTurk) is currently the most popular crowdsourcing platform for research purposes. Researchers ('requesters' in MTurk terminology) upload tasks to MTurk, to be completed by respondents (called 'workers').

Several studies have compared the demographic profile of crowdsourced samples with that of laboratory/university student samples and samples collected over the Internet. Results showed that MTurk workers are more diverse and at least as (if not more) representative of the general population (e.g., [4–9]). The test-retest reliability of MTurk samples has been reported to exceed that of traditionally collected samples [5,10,11]. Other studies have compared the self-reported demographics with criterion data. Rand [12], for example, compared the self-reported country of residence with information extracted from the IP addresses of 176 MTurk workers and found an agreement of 97.2%. Shapiro et al. [13] found that, of the 530 MTurk workers they recruited, 6% reported faulty demographic information. The level of financial compensation does not seem to affect the quality of response, but higher compensation fees have been associated with lower dropout rates [3,14]. Crump et al. [14] replicated 10 classic psychological tasks, including reaction time experiments, rapid stimulus presentation tasks, and learning tasks. MTurk samples behaved similar to the laboratory samples, with an exception of advanced learning tasks, in which MTurk workers had more difficulty. Repeatedly asking the MTurk workers to study the task instructions until they answered all comprehension questions correctly improved task performance. Indeed, it has been found that with the right tutorial, MTurk workers can learn even complex engineering tasks [15].

1.3. CrowdFlower: A recent alternative to MTurk

In October 2007, MTurk stopped supporting requesters from outside the U.S.. Nowadays, the MTurk website reads: "Requesters must provide a U.S. ACH-enabled bank account and a U.S. billing address". Restrictions also apply with respect to workers. Since late 2012, purportedly in an effort to conform to labor laws, MTurk moved to an invitation-only registration. International workers may still be accepted, but the review criteria are not disclosed. The first author of this paper tried to register as an MTurk worker and was kindly rejected with the following e-mail "We regret to inform you that you will not be permitted to work on Mechanical Turk. Our account review criteria are proprietary and we cannot disclose the reason why an invitation to complete registration has been denied".

Various crowdsourcing platforms other than MTurk exist, such as CrowdFlower, CloudCrowd, ShortTask, and MicroWorkers (for an overview see [16] and www.crowdsortium.org). Opposite to MTurk that has its own workforce, CrowdFlower is an aggregator platform delegating tasks to multiple partner channels (e.g., www.clixsense.com) via which users are recruited. More than 1 billion tasks have been completed via CrowdFlower by 5 million contributors, and an estimated five man-years of work is completed on a daily basis. We also reviewed other crowdsourcing platforms, but most appeared to focus on writing, transcribing, tagging, text editing, and Internet searching. CrowdFlower seemed most suitable for scientific (survey) research.

Despite the existence of alternative crowdsourcing services, MTurk has almost monopolized crowdsourcing for scientific research. Chandler et al. [17] argued that the focus on simple and fast tasks, the screening of workers to preserve quality, and the existing infrastructure of MTurk with respect to payments, are three main reasons why scientists use MTurk for research purposes. We did find a number of early studies mentioning CrowdFlower, but even in these cases, CrowdFlower was used to delegate the job to MTurk workers (e.g., [18–20]). This is not possible anymore, because since December 2013 CrowdFlower does not include MTurk in their partners.

In 2014, the number of studies using CrowdFlower for research purposes has increased. CrowdFlower has recently been used in behavioral and psychological experiments [21–24], (psycho)linguistic experiments [25–27], and for investigating public perceptions [28–30].

1.4. Aim of the present study: to investigate the DBQ-accidents relationship using CrowdFlower

In this study we used CrowdFlower to study the relationship between self-reported violations and accidents. Road traffic accidents are a major public health concern, with traffic violations being an important predictor of accidents [31,32]. Young male drivers are a particularly vulnerable group, being involved in a disproportionate number of accidents and traffic violations [31,33].

A previous meta-analysis on the Driver Behaviour Questionnaire (DBQ) [34] showed that men report more violations than women (r = .15), that older drivers report fewer violations than younger drivers (r = −.22), that people who drive more kilometers per year report more violations (r = .06), and that people who report more violations also report more accidents (r = .13). Due to range restriction, these correlations may be underestimates of the true effects in the population. Illustratively, Martinussen et al. [35] found considerably stronger correlations between the DBQ violations scale and age (r = −.46, N = 3,908) among respondents in a broad age range (18–84 years). Recently, the validity of the DBQ has been debated. Af Wåhlberg et al. [36] argued that the correlation between the DBQ and accidents is “too small to be of any practical or theoretical significance” (p. 93).

We aimed to investigate whether the violations-accidents correlations reported in the literature are replicable in an internationally crowdsourced sample. We also investigated the respondents’ knowledge about automated driving, considering that automated driving systems could be a future remedy for traffic violations and road traffic accidents.

Large national differences exist in accident rates, with some low-income countries exhibiting over 10 times more road-traffic fatalities per inhabitant and over 100 times more road-traffic fatalities per registered motor vehicle than some high-income countries [37]. Previous research on cross-national differences in traffic violations using the DBQ pointed toward distinct driving styles between countries, but only a handful of countries were involved in these studies (Great Britain, Finland, and the Netherlands in [38]; Great Britain, Finland, the Netherlands, Greece, Iran, and Turkey in [39]; Finland, Sweden, Greece, and Turkey in [40]; Qatar and United Arab Emirates in [41]; Qatar, Jordan, Indian subcontinent, and Philippines in [42]). We used CrowdFlower to make an international comparison of self-reported traffic violations. Furthermore, we estimated the correlation coefficient between self-reported and registered accident data available from the World Health Organization (WHO) [37].

2. Method

We created a survey on www.crowdflower.com. The survey asked for age, gender, driving frequency and mileage, traffic violations, accident involvement, and familiarity with automated vehicles. The seven violation items were taken from the DBQ as used by De Winter [43], which in turn was taken from Wells et al. [44]. Table 1 shows the variables collected through the survey.

In the instructions we informed the respondents that the survey would take approximately 3 min. The task expiration time was set at 8 min. In order to gather data from an as large and diverse population as possible, no restrictions with respect to the respondents’ country of residence were set, and ‘Level 1 contributors’ were selected, that is, the lowest of the three available levels, accounting for 60% of CrowdFlower’s monthly completed work. Respondents were not allowed to submit multiple surveys from the same IP address. We offered a payment of $0.10 per respondent for completing the survey. We requested 1,862 surveys, so that the cost was lower than $250.

Means and standard deviations were reported for the 16 variables shown in Table 1. Next, Spearman correlations were calculated between self-reported violations and age, gender, mileage, and accidents. Averages of self-reported accidents and violations were analyzed per country (as provided by CrowdFlower by default, based on the respondents’ IP addresses). Spearman correlations were also calculated between the mean number of self-reported accidents per country and the number of road fatalities per 100,000 inhabitants (based on the regression point-estimate of road traffic deaths for 2010 and the population in 2010 per country, both taken from WHO [37]). Spearman correlations were also calculated between the mean number of self-reported accidents and road fatalities

per 100,000 vehicles (based on the regression point-estimate of road traffic deaths for 2010 and the number of registered motor vehicles in 2010 per country, both from WHO [37]). A strong correlation between self-reported and registered data would be indicative of the validity of self-reported accident data obtained from CrowdFlower.

The research was approved by the Human Research Ethics Committee of Delft University of Technology. The survey instructions informed the respondent of the purpose of the research (i.e., "to examine self-reported driving behaviors in different countries of the world"). The first author provided his e-mail address so that respondents could ask questions (none of the respondents actually sent an e-mail). The survey instructions stated that one has to be at least 18 years old to participate, and to continue only when understanding that participation is voluntarily. Informed consent was obtained via a dedicated survey item asking whether the respondent had read and understood the survey instructions. This research and publication are in compliance with the Terms of Use of CrowdFlower (http://www.crowdflower.com/legal; http://www.crowdflower.com/survey).

3. Results

The results were gathered between 16 June 2014 16:59 and 17 June 2014 13:36 Central European Time. CrowdFlower offers respondents the option to provide satisfaction ratings regarding the completed task. The respondents were generally satisfied with both the task and payment (Overall satisfaction = 4.3/5; Instructions clear = 4.5/5; Ease of job = 4.4/5; Test questions fair = 4.1/5; Pay = 4.1/5).

We excluded respondents who did not respond or filled out "no response" in one or more of the multiple-choice questions (N = 200), or indicated they had not read the instructions (N = 11), or were under 18 (N = 6), or never drive (N = 189), or drive 0 km per year (N = 155). Accordingly, 345 unique respondents were excluded, leaving 1,517 respondents for further analysis. For question 15, concerning the year when most cars will be able to drive in fully automated mode, 161 responses were excluded for reporting a year before 2014 (N = 51) or for giving a textual response (N = 110), such as "no idea"/"not sure"/"don't know" (N = 14). We coded the response "never" and similar responses as 9999 years (N = 41). We also excluded phrases which contained a year complemented with text (e.g., "more than 20 years", "in 10 years", or "2030 or maybe never"), because no specific year could be extracted in these cases (e.g., "more than 20 years" could be any year after the year 2034).

Table 1. Definition of variables extracted from the survey and corresponding mean and standard deviation (*SD*) values.

Variable	Full question as reported in the survey	Our coding	Mean (*SD*)
1. Gender	What is your gender?	1 = Female, 2 = Male	0.68 (0.47)
2. Age	What is your age?	Free textual response	32.95 (11.19)
3. DriveFreq	On average, how often did you drive a vehicle in the last 12 months?	From 0 = Never to 5 = Every day	3.85 (1.18)
4. NrAcc	How many accidents were you involved in when you were driving a car in the last 3 years?	From 0 to 5 accidents; a response "more than 5" was coded as 6	0.43 (0.88)
5. KmYear	About how many kilometers did you drive in the last 12 months? (If you are not certain, please give as good an estimate as you can)	From 0 = 0 km to 10 = More than 100,000 km	3.59 (2.06)
6. Vangered	How often do you do the following?: Becoming angered by a particular type of driver, and indicate your hostility by whatever means you can.	From 0 = Never to 5 = Nearly all the time	1.32 (1.12)
7. Vmotorway	How often do you do the following?: Disregarding the speed limit on a motorway.	From 0 = Never to 5 = Nearly all the time	1.33 (1.24)
8. Vresident	How often do you do the following?: Disregarding the speed limit on a residential road.	From 0 = Never to 5 = Nearly all the time	1.36 (1.22)
9. Vfollowing	How often do you do the following?: Driving so close to the car in front that it would be difficult to stop in an emergency.	From 0 = Never to 5 = Nearly all the time	0.75 (0.98)
10. Vrace	How often do you do the following?: Racing away from traffic lights with the intention of beating the driver next to you.	From 0 = Never to 5 = Nearly all the time	0.55 (0.95)
11. Vhorn	How often do you do the following?: Sounding your horn to indicate your annoyance with another road user.	From 0 = Never to 5 = Nearly all the time	1.42 (1.06)
12. Vphone	How often do you do the following?: Using a mobile phone without a hands-free kit.	From 0 = Never to 5 = Nearly all the time	0.98 (1.21)
13. Vmean	–	Mean across the above 7 violation items	1.10 (0.71)
14. Google	Have you heard of the Google Driverless Car?	0 = No, 1 = Yes	0.50 (0.50)
15. YearAuto	In which year you think that most cars will be able to drive fully automatically on the roads in your country?	Free textual response	2030[a] (1028)
16. SurvTime	–	Survey completion time in seconds	182.4 (94.2)

[a]For YearAuto, the median is reported instead of the mean because the distribution was highly skewed.

3.1. Analysis at the individual level

Table 1 shows the descriptive statistics. The respondents' mean age was 32.95 years, and 68% were male. The respondents reported on average 0.43 accidents in the past three years. Becoming angered at another driver, speeding (on a motorway or residential road), and horn honking were the most common violations, whereas racing away from a traffic light was the least common one. Half of the respondents had heard of the Google Driverless Car before. People expected most cars to be able to drive in fully automatically around 2030 (median value).

Table 2 shows the correlation matrix of the variables. Consistent with the meta-analysis by De Winter and Dodou [34], men reported more violations than women (ρ = .19), older persons reported fewer violations than younger persons (ρ = −.10), people with a higher driving frequency reported more violations (ρ = .12), and people who drove more kilometers per year reported more violations (ρ = .18). The number of violations correlated quite strongly with self-reported accidents (ρ = .28), but not with familiarity with the Google Driverless Car (ρ = −.01) or the expected year of fully automated cars on the road (ρ = .00).

3.2. Analysis at the national level

The 1,517 respondents came from 88 different countries. To prevent erratic effects due to sampling error we selected only the 18 countries with 25 or more respondents. Table 3 shows clear national differences in self-reported accident rates and violations. Respondents in India and Indonesia reported the largest accident rates. Sounding a horn to indicate annoyance also seemed common in these two countries. Among Western countries, respondents in Spain and Italy reported more horn honking than respondents in Germany, in line with Forgas [45], who in a field experiment found that drivers in Spain and Italy (and France) honk faster than drivers in Germany. In some Western countries (e.g., Canada, Germany, Poland, Portugal), it was relatively common to disregard the speed limit.

We calculated the correlation between self-reported accidents per country and the annual number of fatal accidents per 100,000 inhabitants according to WHO [37]. The results in Figure 1 reveal a strong association (ρ = .79, $p < .001$, $N = 18$). The correlation between self-reported accidents and the number of road fatalities per 100,000 vehicles was similar (ρ = .68, $p = .002$, $N = 18$). The correlation between the countries' mean violation score and the annual number of fatal accidents per inhabitant was ρ = .58 ($p = .013$; $N = 18$), whereas the correlation between the countries' mean violation score and the number of road fatalities per 100,000 vehicles was ρ = .47 ($p = .053$; $N = 18$). The correlations between self-reported violations and accidents within each country were overall positive, with a sample-size weighted correlation of ρ = .26 (Table 3).

Very fast respondents (the fastest 5%, $M = 55.9$ s, $SD = 9.7$ s, $N = 75$) showed the strongest Vmean-NrAcc correlation (ρ = .55). Finally, we observed that respondents from English-speaking countries (Canada, U.S., Great Britain) took shorter time to fill out the survey than respondents from non-English-speaking countries (Table 3).

Table 2. Spearman correlation matrix at individual respondent level ($N = 1,517$).

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1.Gender															
2.Age	−.22														
3.DriveFreq	.05	.17													
4.NrAcc	.13	−.14	.12												
5.KmYear	.05	.21	.50	.09											
6.Vangered	.11	−.09	.05	.19	.08										
7.Vmotorway	.09	.08	.07	.09	.14	.25									
8.Vresident	.10	−.03	.09	.16	.12	.29	.54								
9.Vfollowing	.11	−.12	.01	.21	.04	.25	.28	.32							
10.Vrace	.12	−.08	.05	.14	.11	.35	.32	.37	.35						
11.Vhorn	.15	−.13	.12	.28	.11	.37	.17	.30	.20	.26					
12.Vphone	.12	−.09	.11	.22	.18	.25	.25	.30	.29	.29	.25				
13.Vmean	.19	−.10	.12	.28	.18	.61	.65	.72	.56	.59	.57	.58			
14.Google	.10	.01	−.01	.03	.02	−.01	.07	−.01	−.01	.04	−.04	−.07	−.01		
15.YearAuto	.04	.00	−.05	−.04	−.04	.02	−.01	−.01	.00	−.03	−.04	.02	.00	−.14	
16.SurvTime	.08	.02	−.06	.10	−.10	.02	−.02	−.01	.05	−.08	.12	.04	.05	−.09	.11

Table 3. Means per country for the 16 variables defined in Table 1 and Spearman (ρ) and Pearson (*r*) correlations between Vmean and NrAcc.

	Respond-ents	1.Gender	2.Age	3.Drive-Freq	4.NrAcc	5.KmYear	6.Vangered	7.Vmotor way	8.Vresident	9.Vfollow-ing	10.Vrace	11.Vhorn	12.Vphone	13.Vmean	14.Google	15.Year-Auto[a]	16.Surv-Time	ρ Vmean-NrAcc	r Vmean-NrAcc
BGR	30	0.60	35.5	3.43	0.17	2.57	1.03	0.93	0.93	0.87	0.30	1.17	1.07	0.90	0.27	2050	234	.35	.31
BRA	30	0.77	27.1	3.77	0.50	3.73	1.53	1.40	1.27	0.97	0.23	1.17	0.67	1.03	0.50	2040	213	.39	.39
CAN	108	0.42	40.3	3.91	0.19	3.85	1.10	1.68	1.47	0.50	0.42	1.09	0.46	0.96	0.48	2025	151	.19	.19
DEU	26	0.73	36.5	3.88	0.15	3.96	1.19	1.58	1.77	0.81	0.46	0.88	0.54	1.03	0.54	2035	177	.49	.57
ESP	65	0.72	30.2	3.85	0.29	3.46	1.22	1.29	1.20	0.72	0.42	1.40	0.71	0.99	0.57	2030	162	.47	.53
GBR	107	0.51	38.7	3.99	0.12	3.64	1.15	1.30	1.08	0.39	0.50	1.08	0.26	0.82	0.58	2030	124	.08	.07
GRC	39	0.67	32.5	4.31	0.41	3.85	1.46	1.49	1.28	0.79	0.44	1.23	1.23	1.13	0.49	2050	195	.15	.15
IDN	51	0.78	28.9	3.94	0.71	3.59	1.49	0.80	1.10	0.98	0.67	1.80	1.31	1.17	0.51	2030	235	.46	.35
IND	144	0.84	28.2	4.15	0.83	3.14	1.55	1.25	1.44	0.93	0.74	2.28	1.19	1.34	0.53	2030	203	.26	.28
ITA	41	0.73	34.2	3.95	0.37	4.22	1.34	1.44	1.49	0.71	0.51	1.32	0.71	1.07	0.41	2030	178	.38	.33
MYS	25	0.56	28.7	4.28	0.56	4.12	1.60	1.04	1.44	1.08	0.68	1.20	1.52	1.22	0.40	2030	209	.30	.31
PHL	50	0.58	30.3	3.14	0.34	2.96	1.10	0.74	0.82	0.66	0.46	1.58	1.02	0.91	0.42	2025	186	.14	.14
POL	25	0.72	29.6	3.96	0.56	4.04	1.60	1.44	1.88	0.48	0.64	0.96	0.92	1.13	0.56	2030	166	.09	−.01
PRT	49	0.71	30.3	4.35	0.35	4.08	1.31	1.86	1.88	0.63	0.35	1.16	0.96	1.16	0.51	2030	178	.19	.12
ROU	55	0.76	31.9	3.75	0.29	3.33	1.42	1.05	1.78	0.84	0.33	1.56	1.20	1.17	0.49	2050	186	.51	.55
SRB	26	0.73	31.7	3.19	0.38	2.31	1.62	1.62	1.38	0.46	0.38	1.50	1.23	1.17	0.58	2050	210	.59	.62
USA	141	0.39	40.2	4.05	0.16	4.16	0.93	1.60	1.35	0.49	0.55	1.09	0.84	0.98	0.54	2030	133	.22	.31
VEN	32	0.91	28.2	3.97	0.53	3.47	1.09	1.22	1.19	0.56	0.47	1.75	1.31	1.08	0.31	2050	225	−.10	.07

[a]For YearAuto, the median is reported instead of the mean because the distribution was highly skewed (25.3).
Country abbreviations according to ISO 3166-1 alpha-3.

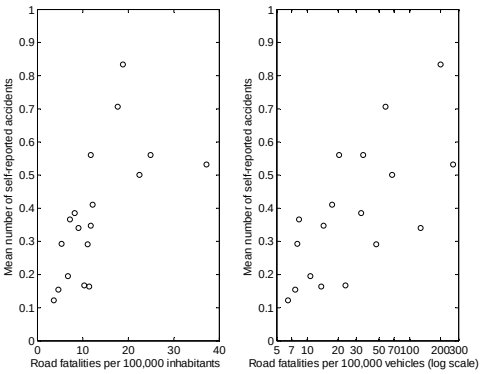

Fig. 1. Self-reported accidents versus road fatality statistics according to WHO (only the 18 countries having at least 25 respondents are shown).

4. Discussion

CrowdFlower turned out to be an efficient tool for gathering survey data. For $247 we obtained 1,862 responses within 20 hours. This result is in line with Crump et al. [14] who replicated various classical psychological tasks within a few hours via MTurk and concluded that "it was amazing how much data we could collect in a short period of time" (p. 16). As a reference, we requested a quote from a Dutch marketing agent, which had been used before by colleagues at our institution for a traffic psychology survey. For conducting a national Internet survey using 15 questions and 2,000 respondents, the marketing agent requested 7,495 € ($9,097) excluding VAT.

Correlations between the DBQ and age, gender, driving exposure, and self-reported accidents had the same direction as those in the literature [34]. At the national level, self-reported accidents and objective road safety data correlated fairly strongly (ρ = .68/.79, N = 18), suggesting that the self-reported accident data were valid.

The vast majority of DBQ research (i.e., several hundreds of studies by now [34,46]) has been conducted within a single country, whereas the most cross-cultural DBQ study so far was conducted across six countries [39]. Our sample contained 18 countries with 25 or more respondents, making it the most international DBQ sample to date.

The correlation between violations and self-reported accidents (ρ = .28 [r = .29]) was considerably stronger compared to past DBQ research (r = .13 in a meta-analysis of 23 samples [34]; r = .14 in a meta-analysis of 67 samples [46]; r = .19 in a study using the exact same violations scale [43]). One possible explanation is common method variance, as suggested by the violations-accidents correlation being strongest among the fastest respondents.

Another explanation is the large variance of the predictor and criterion variables. Our sample included both countries with a very high number of road accidents (e.g., Malaysia, Brazil, India, Indonesia) and countries with a very low number of accidents (e.g., Great Britain, Germany). Indeed, the correlation between violations and self-reported accidents was on average (slightly) weaker ($\rho = .26$) per country than across countries (Table 3).

CrowdFlower provides mechanisms to prevent contamination from scammers who try to abuse the crowdsourcing services for earning as much money as possible. Specifically, a requester can reject the work such that the worker does not get paid, and a quality control service can be applied which requires workers to answer test questions. We have not made use of these options. However, from the 1,862 respondents, only 11 respondents (0.6%) indicated they had not read the instructions, 21 did not fill in their age, and 2 submitted an unrealistic age (i.e., an age of 3 and 222, respectively). Furthermore, respondents, on average, took a reasonable amount of time (3 min) to complete the survey, which corresponds to the duration mentioned in the survey instructions.

Some have argued that crowdsourced samples are more diverse than laboratory samples, while others have warned that crowdsourced samples are generally younger [13], more highly educated [47,48], yet underemployed [8,13], compared to the general population. According to CrowdFlower, half of their workers have an annual household income of less than $10,000 (http://www.crowdflower.com/blog/2014/01/demographics-of-the-largest-on-demand-workforce). Differences in Internet access and financial incentives may cause bias among countries.

Another concern regarding the external validity of crowdsourced samples is that some respondents may have evolved into “professional” crowd workers [17]. Paolacci and Chandler [49] reported that 10% of the workers appeared to be responsible for completing 41% of all tasks on MTurk. This means that some workers gain high experience in tasks, thereby becoming atypical of the general population. It has been also reported that some workers follow their favorite requesters, which creates a dependency between results generated by the same research group [17,50]. Such concerns do not seem to apply to our study, as the use of CrowdFlower for scientific surveys is relatively new.

Finally, our questions on automated driving are useful for shaping hypotheses about remedial measures for traffic violations. In our survey 74% of the respondents argued that fully automated driving will reach a 50% market share between now and 2050. In 2014, a survey among automotive experts showed that about 35% and 30% of them expected fully automated vehicles driving on the roads before 2030 and 2040, respectively [51]. On the other hand, we found it surprising that only half of the CrowdFlower respondents had heard of the Google Driverless Car, even in the U.S.. Given the media attention surrounding automated driving, we had expected this percentage to be higher.

Acknowledgements

JW, MK, and RH are involved in the Marie Curie ITN: HFAuto (PITN-GA-2013-605817).

References

[1] Gosling SD et al. (2004) Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *Am Psychol 59*, 93–104.

[2] Rentfrow PJ et al. (2013) Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *J Pers Soc Psychol 105*, 996–1012.

[3] Mason W, Suri S (2012) Conducting behavioral research on Amazon’s Mechanical Turk. *Behav Res Methods 44*, 1–23.

[4] Berinsky AJ et al. (2012) Evaluating online labor markets for experimental research: Amazon. com’s Mechanical Turk. *Polit Anal 20*, 351–368.

[5] Buhrmester M et al. (2011) Amazon’s Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspect Psychol Sci 6*, 3–5.

[6] Goodman JK et al. (2013) Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J Behav Decis Mak 26*, 213–224.

[7] Horton JJ et al. (2011) The online laboratory: Conducting experiments in a real labor market. *Exp Econ 14*, 399–425.

[8] Paolacci G et al. (2010) Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak 5*, 411–419.

[9] Simons DJ, Chabris CF (2012) Common (mis) beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PLOS ONE 7*, e51876.

[10] Behrend TS et al. (2011) The viability of crowdsourcing for survey research. *Behav Res Methods 43*, 800–813.

[11] Holden CJ et al. (2013) Assessing the reliability of the M5-120 on Amazon’s Mechanical Turk. *Comput Human Behav 29*, 1749–1754.

[12] Rand DG (2012) The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *J Theor Biol 299*, 172–179.
[13] Shapiro DN et al. (2013) Using Mechanical Turk to study clinical populations. *Clin Psychol Sci 1*, 213–220.
[14] Crump MJ et al. (2013) Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE 8*, e57410.
[15] Staffelbach M et al. (2014) Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with Amazon Mechanical Turk. *Collective Intelligence Conference*, Cambridge, MA.
[16] Vakharia D, Lease M (2013) *Beyond AMT: An analysis of crowd work platforms.* Retrieved from arxiv.org/abs/1310.1672v1.pdf.
[17] Chandler J et al. (2014) Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behav Res Methods 46*, 112–130.
[18] Smucker MD, Jethani CP (2011) The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior. *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*.
[19] Le J et al. (2010) Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation* (pp 21–26).
[20] Negri M, Mehdad Y (2010) Creating a bi-lingual entailment corpus through translations with Mechanical Turk: $100 for a 10-day rush. *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp 212–216).
[21] Hindriks F et al. (2014) A new angle on the knobe effect: Intentionality correlates with blame, not with praise. Retrieved from http://singmann.org/download/publications/submitted/Knobe-correlations2_H_word.pdf.
[22] Nguyen D et al. (2014) Using crowdsourcing to investigate perception of narrative similarity. *23rd ACM International Conference on Information and Knowledge Management* (pp 321–330).
[23] Pepper GV, Nettle D (2014) Out of control mortality matters: The effect of perceived uncontrollable mortality risk on a health-related decision. *PeerJ 2*, e459.
[24] Wolf W et al. (2015) Ostracism Online: A social media ostracism paradigm. *Behav Res Methods 47*, 361–373.
[25] Marelli M et al. (in press) Picking buttercups and eating butter cups: Spelling alternations, semantic relatedness, and their consequences for compound processing. *Appl Psycholinguist*.
[26] Over D et al. (2013) Scope ambiguities and conditionals. *Think Reason 19*, 284–307.
[27] Wang S et al. (2014) Exploring mental lexicon in an efficient and economic way: Crowdsourcing method for linguistic experiments. *COLING 2014* (pp 105–113).
[28] Aladhadh S et al. (2014) Tweet author location impacts on tweet credibility. *Australasian Document Computing Symp*.
[29] Nawrot I, Doucet A (2014) Timeline localization. In *Human-computer interaction. Theories, methods, and tools* (pp 611–622) Springer.
[30] Prpić J et al. (2014) Experiments on crowdsourcing policy assessment. *Ann Rev Policy Design, 2*, 1–10.
[31] Evans L (2004) *Traffic safety.* Bloomfield Hills, MI: Science Serving Society.
[32] Redelmeier DA, McLellan BA (2013) Modern medicine is neglecting road traffic crashes. *PLOS Med 10*, e1001463.
[33] Lee JD (2007) Technology and teen drivers. *J Safety Res 38*, 203–213.
[34] De Winter JCF, Dodou D (2010) The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis. *J Safety Res 41*, 463–470.
[35] Martinussen LM et al. (2014) Assessing the relationship between the Driver Behavior Questionnaire and the Driver Skill Inventory: Revealing sub-groups of drivers. *Transp Res Part F Traffic Psychol Behav 26*, 82–91.
[36] Af Wåhlberg A et al. (2012) Commentary on the rebuttal by de Winter and Dodou. *J Safety Res 43*, 90–93.
[37] World Health Organization (2013) WHO global status report on road safety 2013: supporting a decade of action. WHO.
[38] Lajunen T et al. (2004) The Manchester driver behaviour questionnaire: a cross-cultural study. *Accid Anal Prev 36*, 231–238.
[39] Özkan T et al. (2006) Cross-cultural differences in driving behaviours: A comparison of six countries. *Transp Res Part F Traffic Psychol Behav 9*, 227–242.
[40] Warner HW et al. (2011) Cross-cultural comparison of drivers' tendency to commit different aberrant driving behaviours. *Transp Res Part F Traffic Psychol Behav 14*, 390–399.
[41] Bener A et al. (2008) The driver behaviour questionnaire in arab gulf countries: Qatar and united arab emirates. *Accid Anal Prev 40*, 1411–1417.
[42] Bener A et al. (2013) A cross "ethnical" comparison of the Driver Behaviour Questionnaire (DBQ) in an economically fast developing country. *Glob J Health Sci 5*, 165–175.
[43] De Winter JCF (2013) Predicting self-reported violations among novice license drivers using pre-license simulator measures. *Accid Anal Prev 52*, 71–79.
[44] Wells P et al. (2008) Cohort II: A study of learner and new drivers. *Volume 1 - Main report*. London: Dep Transport.
[45] Forgas JP (1976) An unobtrusive study of reactions to national stereotypes in four European countries. *J Soc Psychol 99*, 37–42.
[46] De Winter JCF et al. (in press) A quarter of a century of the DBQ: some supplementary notes on its validity with regard to accidents. *Ergonomics*.
[47] Cooper EA, Farid H (in press) Does the Sun revolve around the Earth? A comparison between the general public and online survey respondents in basic scientific knowledge. *Public Underst Sci*.
[48] Kang R et al. (2014) Privacy attitudes of mechanical turk workers and the US public. *Symposium on Usable Privacy and Security*.
[49] Paolacci G, Chandler J (2014) Inside the Turk. Understanding Mechanical Turk as a participant pool. *Curr Dir Psychol Sci 23*, 184–188.
[50] Stewart N et al. (2015) *The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers.* Retrieved from https://www.stewart.warwick.ac.uk/publications/MTurk.html
[51] Underwood SE (2014) Automated vehicles forecast. Vehicle Symposium Opinion Survey. *Automated Vehicles Symp, San Francisco*.